Endpoint configurations

The following are configurable settings within an Endpoint.

Endpoint Name

Create a name you'd like to use for the Endpoint configuration. The resulting endpoint will be assigned a random ID to be used when making calls.

The name is only visible to you.

GPU Selection

Select one or more GPUs you want your Endpoint to run on.

Active (Min) Workers

Setting this amount to 1 will result in "always on" workers. This will allow you to have a worker ready to respond to job requests without incurring any cold start delay.

> **NOTE**
>
> You will incur the cost of any active workers you have set regardless if they are working on a job.

Max Workers

This will establish a ceiling or upper limit to the number of active workers your endpoint will have running at any given point.

Default: 3

▸ How to configure Max Workers

GPUs / Worker

The number of GPUs you would like assigned to your worker.

> ⓘ **NOTE**
>
> Currently only available for 48GB GPUs.

Idle Timeout

The amount of time in seconds a worker not currently processing a job will remain active until it is put back into standby. During the idle period, your worker is considered running and will incur a charge.

Default: 5 seconds

FlashBoot

RunPod magic to further reduce the average cold-start time of your endpoint. FlashBoot works best when an endpoint receives consistent utilization. There is no additional cost associated with FlashBoot.

Advanced

Additional controls to help you control where your endpoint is deployed and how it responds to incoming requests.

Data Centers

Control which datacenters you would like your workers deployed and cached. By default all datacenters are selected.

Select Network Volume

Attach a network storage volume to your deployed workers.

Ask AI

Network volumes will be mounted to `/runpod-volume/`.

> **NOTE**
>
> While this is a high performance network drive, do keep in mind that it will have higher latency than a local drive.
>
> This will limit the availability of cards, as your endpoint workers will be locked to the datacenter that houses your network volume.

Scale Type

- **Queue Delay** scaling strategy adjusts worker numbers based on request wait times. With zero workers initially, the first request adds one worker. Subsequent requests add workers only after waiting in the queue for the defined number of delay seconds.
- **Request Count** scaling strategy adjusts worker numbers according to total requests in the queue and in progress. It automatically adds workers as the number of requests increases, ensuring tasks are handled efficiently.

```
_Total Workers Formula: Math.ceil((requestsInQueue + requestsInProgress)
```

GPU Types

Within the select GPU size category you can further select which GPU models you would like your endpoint workers to run on. Default: `4090` | `A4000` | `A4500`

▸ What's the difference between GPU models.

Edit this page

Previous
« References

Next
Endpoint operations »

Ask AI

Docs

Overview

Tutorials

AI APIs

Community

Discord

Contact us

More

Blog

GitHub

Ask AI